蒙日機械翻訳システムー訳語選択を中心としてー

江原暉将　　早田清冷　　木村展幸

諏訪東京理科大学　　東京外国語大学　　(株)漢字情報サービス
eharate@rs.suwa.tus.ac.jp　　saksaha@spn1.speednet.ne.jp　　kimura@kiss.co.jp
http://www.rs.suwa.tus.ac.jp/eharate

1　はじめに

モンゴル語と日本語は語順が類似していることを活用して蒙日機械翻訳システムを研究している。[江原ほか 2004]では、モンゴル語の形態素解析について、[江原ほか 2005]では蒙日辞書を用いた日本語単語列の出力について述べた。本文では、解析・変換・生成から成るシステムの全体構成について述べると共に、最も重要な課題である語彙変換(訳語選択)の手法について詳述する。

モンゴル語の計算機処理に関する研究には[那順ほか 2001][満ほか 2005][Enkhbayar ほか 2005]などがあるが、蒙日機械翻訳を扱ったものは筆者の知る限り存在しない。

2　システム構成

本システムは伝統的なトランスファー方式であり、解析、変換、生成の 3 つのフェーズから成る。

2.1　解析フェーズ

解析フェーズのうち形態素解析は処理系に茶筌を用い、モンゴル語の辞書や文法表、連接規則などを作成して構築した。モンゴル語の構文解析は行っていない。モンゴル語と日本語は語順が極めて類似しており[江原 1995]、後述する構造変換(語順変更)は局所的と成る。そのためモンゴル語の構文情報は必要としないためである[1]。

2.2　変換フェーズ

変換フェーズは一般的に構造変換と語彙変換から成る[江原、田中 1993]。ラベル付き有向グラフ[2]として表現された解析結果の構造を変更するのが構造変換であり、ラベルを変更するのが語彙変換である。

本システムの場合、構造変換は局所的であり、形態素列上でパターンとアクションの組から成る変換規則を用いて行っている。そのための処理系は[江原ほか 2000]と同様のものを用いている。構造変換の規則の例を図 1 に示す。モンゴル語と日本語では、動詞、否定の助辞、過去の助辞の語順が異なるので、語順を変更する規則である。これによって「知らなかった」という日本語の語順が得られる。現在用いている構造変換規則の数は 7 である。

規則
動詞+過去+否定 ⇒ 動詞+否定+過去

実行例
м э д(知る)＋с э н(過去)＋г ү й(否定)
⇒　知る+否定+過去

図 1　構造変換規則の例

語彙変換は 3 つの部分に分けられる。(1)複合語の変換、(2)空要素の補完、(3)訳語選択、である。このうち(1)と(2)は構造変換と同様、規則を用いて処理している。(3)については節を改めて説明する。

複合語の変換は、逐語訳では問題がある場合に、複合語(語列)パターンを用いて訳出しする規則で図 2 に例を示す。本例は、通常「て」または「で」と訳されるжをэ х л э хに前接する場合は、空要素にする規則である。これによって例えば「行ってはじめる」ではなく「行きはじめる」という日本語が得られる。複合語の変換規則は、現在 8 ある。

規則
動詞+ж(て)+э х л э х(はじめる)
⇒ 動詞+はじめる

図 2　複合語変換規則の例

---

[1] 語彙変換(訳語選択)のためにモンゴル語の構文情報を利用することはもちろん考えられるが、ここでは利用していない。

[2] 一般的には木で表現されることが多いが、本システムの場合は形態素列つまり線状のグラフである。

空要素の補完は、モンゴル語では空であるが日本語では語が必要な場合に用いられる規則であり、図 3 に例を示す。本例は、モンゴル語で格語尾が付加していない名詞に日本語で必要な格助詞を補完する規則である。補完する格助詞にはあいまい性がある。あいまい性の解消は(3)の訳語選択処理で統一的に行っている。空要素の補完規則は現在 15 ある。

規則
名詞+ゼロ格語尾
⇒ 名詞+{〈格助詞〉/【】/【の】/【と】/【が】/【は】/【を】/【に】/【で】}

図 3　空要素の補完規則の例

2.3　生成フェーズ

生成フェーズでは、変換フェーズの結果得られた日本語単語列(標準形で表記されている)に適切な活用処理を加えて日本語文を出力する。まず、活用語に対して全ての活用形に展開し、(日本語の)連接規則を用いて連接可能な活用形を選択する。

以上述べた本システムの全体構成を通常の構文トランスファー方式と比較して図 4 に示す。通常のシステムでは構文解析と構造変換が必要だが、本システムでは構文解析はなく、構造変換も小さなものである。

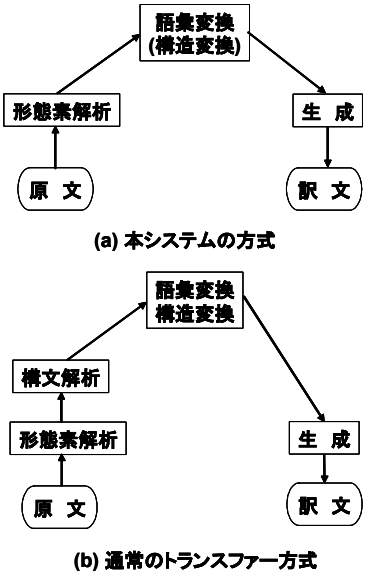

図 4　全体構成

3　訳語選択

前節で述べたように本システムでは訳語選択が中心的課題となる。本システムで用いている訳語選択の手法は確率モデルによるものである。一般的に確率モデルによる機械翻訳は以下のように定式化される。入力モンゴル語文を m とし、日本語の文の全体を J とするとき、機械翻訳結果 $\hat{j}$ は m が与えられたときの事後確率 $P(j|m)$ を最大にするものとして

$$\hat{j} = \arg\max_{j \in J} P(j|m) \qquad (1)$$

で与えられる[Brown et al. 1990]。通常は、(1)を Bayes の定理を用いて変形するが、ここでは、異なるやり方で変形する。(1)の右辺に $1 = P(j)/P(j)$ を乗じて

$$\hat{j} = \arg\max_{j \in J} P(j) \times P(j|m)/P(j) \qquad (2)$$

とする。$P(j)$ を言語モデルと呼び、$P(j|m)/P(j)$ を翻訳モデルと呼ぶ。

一方、本システムでは、構造変換や空要素の処理は規則方式を用いて終了しているので、確率モデルによる訳語選択では、m と j の語数は同一であり、語順も一致している。そこで(2)は以下のように変形できる。m を $m = m_1, m_2, ..., m_n$ と単語列[3]に分解し、各語 $m_i (i = 1, ..., n)$ の日本語訳語の集合を J_i とすると(2)は

$$\hat{j} = \arg\max_{j \in J_1 \times \cdots \times J_n} P(j) \times P(j|m)/P(j) \qquad (3)$$

とできる。さらに翻訳モデルに単語生起の独立性を仮定すると(3)は

$$\hat{j} = \arg\max_{(j_1, \cdots j_n) \in J_1 \times \cdots \times J_n} \{P(j_1 \cdots j_n) \times \prod_{i=1}^{n} P(j_i|m_i)/P(j_i)\} \qquad (4)$$

となる。つまり翻訳モデルは日本語単語 j_i の事前確率とモンゴル単語 m_i が現れた時の事後確率の比を単語数 n にわたって掛けた値となる。$P(j_i|m_i)$ は、事後確率が付与された蒙日辞書があれば求められるが、我々は、現時点でそのような辞書を持っていない。そこで

[3] 正確には形態素列。

$$P(j_i \mid m_i) \propto \begin{cases} P(j_i) & (j_i \in J_i) \\ 0 & (else) \end{cases} \quad (5)$$

と仮定する。すると(4)は

$$\hat{j} = \underset{(j_1,\cdots j_n)\in J_1\times\cdots\times J_n}{\arg\max} P(j_1 \cdots j_n) \quad (6)$$

と簡単になる。翻訳モデルは訳文を訳語の直積集合の要素に制約することにのみ用いられ、確率は言語モデルのみで計算される。

4　言語モデル

日本語単語列 $j = j_1, j_2, \cdots, j_n$ の事前確率を計算する言語モデルとしては、2 種のものを考える。文節バイグラムモデルと係り受けモデルである。係り受けモデルは文節列を対象にしているので、バイグラムモデルも文節を単位とする。そこで単語列の代わりに文節列を用いるが、形式的には同じであるので $j = j_1, j_2, \cdots, j_n$ を文節列とする。文節バイグラムモデルは

$$P(j_1, j_2, \cdots, j_n) = P(j_n)\prod_{i=1}^{n-1} P(j_i \mid j_{i+1}) \quad (7)$$

とするもので、係り受けモデルは

$$P(j_1, j_2, \cdots, j_n) = P(j_n)\prod_{i=1}^{n-1} P(j_i \mid j_{k_i}) \quad (8)$$

となる。ここで、文節 j_{ki} は文節 j_i の係り先であり、(8)を最大にする係り先を選ぶ。実際には、最後の文節 j_n を"EOS"としたので、(7)や(8)の $P(j_n)$ は無視してよい。

言語モデル確率は、毎日新聞の 1991 年、1992 年および 1994 年から 2000 年までの 9 年分の記事[毎日新聞社]から推定した。総文数は約 890 万文である。これらの文を茶筌と南瓜を用いて形態素解析と係り受け解析を行い、データ数をカウントして(7)(8)の条件付確率を求めた[4]。

ここで、文節を内容語部分と機能語部分に分け、バイグラムモデルの確率計算においては、後文節 j_{i+1} の機能語部分は無視し、係り受けモデルの確率計算においては係り先文節 j_{ki} の機能語部分は無視した。

5　翻訳実験

前節までで述べたシステムを用いて翻訳実験を行った。用いた蒙日辞書は[江原 2005]と同一のものである。実験用のモンゴル語原文は[小沢 1986]の練習問題から 182 文を抽出して用いた。本文献はシステム開発のときに参考にした文献であり、実験は close test となる。また、実験用の文は教科書レベルのものであり、1 文あたりの平均語数は 5.2 と短い。モンゴル語の 1 形態素あたりの日本語訳語数の分布を図 5 に示す。訳語数が 5 と 8 の部分で度数が多いが、これはモンゴル語で空である格要素に日本語で複数の助詞を補完したためである。

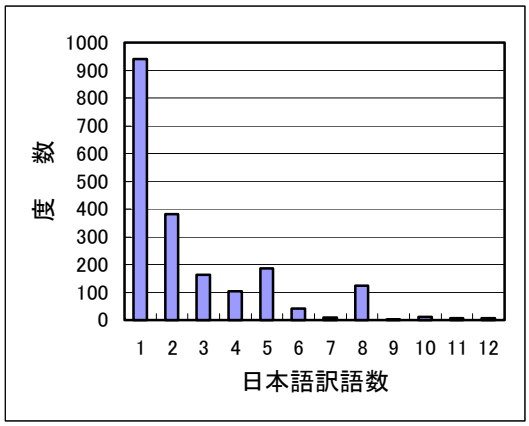

図 5　日本語訳語数の分布

結果を形態素単位の BLEU[5]で評価したものを表 1 に示す。言語モデルとして係り受けモデルを用いた方がバイグラムモデルを用いるより精度が高かった。

表 1　実験結果

言語モデル	BLEU
係り受け	0.217
バイグラム	0.194

表 2 に翻訳結果の例を示す。訳語選択の誤りが多いが形態素解析の誤りもある。

6　おわりに

蒙日機械翻訳システムについて、訳語選択を中心に述べた。今後の課題としては以下のようなことが考えられる。

対訳辞書の充実：辞書の語数を増やすことはもちろんであるが、訳語選択能力を強化するためには、事後確率が付与された対訳辞書の構築

[4] バージョンは以下の通り。茶筌：2.3.3、ipadic：2.7.0、南瓜：0.52。

[5] 日本語の形態素を単位とし 4-gram の累積値である。計算には ftp://jaguar.ncsl.nist.gov/mt/resources/の中の mteval-v11a.pl を利用した。

が望まれる。事後確率は単語対単語の対応だけでなく、モンゴル単語の周辺文脈に依存した確率であってもよい。

対訳コーパスの構築：翻訳精度を一層向上させるためには対訳コーパスを構築し、そこから翻訳知識を抽出する必要がある。

他言語への展開：日本語と語順が類似した言語は他にもある。それらの言語に本手法を展開することが考えられる。

参考文献

[Brown et al. 1990] Peter F. Brown; John Cocke; Stephen A. Della Pietra; Vincent J. Della Pietra; Fredrick Jelinek; John D. Lafferty; Robert L. Mercer; Paul S. Roossin: A statistical Approach to Machine Translation, Computational Linguistics, Volume 16, Number 2, pp.79-85, June 1990.

[江原、田中 1993] 江原暉将、田中穂積： 機械翻訳における自然言語処理、情報処理、自然言語処理技術の応用特集、Vol.34, No.10, pp.1266-1273, Oct. 1993.

[江原 1995] 江原暉将：多次元尺度構成法を用いた語順パラメータの間の関係付け、言語処理学会第 1 回年次大会発表論文集、pp.173-176, Mar. 1995.

[江原ほか 2000] 江原暉将、福島孝博、和田裕二、白井克彦：聴覚障害者向け字幕放送のためのニュース文自動短文分割、電子情報通信学会技術研究報告、NLC2000-12, pp.17-22, July 2000..

[江原ほか 2004] 江原暉将、早田清冷、木村展幸：茶筌を用いたモンゴル語の形態素解析、言語処理学会第 10 回年次大会発表論文集、pp.709-712, Mar. 2004.

[江原ほか 2005] 江原暉将、早田清冷、木村展幸：茶筌を用いたモンゴル語から日本語への機械翻訳、言語処理学会第 11 回年次大会発表論文集、pp.534-537, Mar. 2005.

[Enkhbayar ほか 2005] Sanduijav Enkhbayar、宇津呂武仁、佐藤理史：音韻論的・形態論的制約を用いたモンゴル語句生成・形態素解析、自然言語処理、Vol.12, No.5, pp.185-206, Oct.2005.

[毎日新聞社] 毎日新聞社：CD 毎日新聞(データ集)、1991 年版, 1992 年版, 1994 年版～2000 年版.

[満ほか 2005] 満都拉, 藤井敦, 石川徹也：伝統的モンゴル語の電子化方式とテキスト検索への応用、電子情報通信学会論文誌、Vol.J88-D2, No.10, pp.2102-2111, Oct. 2005.

[那順ほか 2001] 那順烏日图、刘群、巴达玛敷德則尔：面向机器翻訳的蒙古語生成、自然語言理解与机器翻訳、精華大学出版社、June 2001.

[小沢 1986] 小沢重男：モンゴール語四週間、大学書林, 1961.

表 2　蒙日翻訳例

No.	モンゴル語原文	蒙日翻訳結果	基準日本語訳文
4	Өвс, гэрлийн хүчээр ургана.	草，明かりの力で実る.	草は光のおがげで育つ。
5	Өвс чийгийн хүчээр ургана.	草湿気の力で実る.	草は湿気の力で育つ。
6	Өвс газрын шимээр ургана.	草が大地の栄養で生える.	草は土地の栄養によって育つ。
7	Мал өвсөөр амьдарна.	家畜草で暮らす.	家畜は草によって生活する。
8	Монгол хүн, малаар амьдарна.	モンゴルの人が，家畜で暮らす.	モンゴル人は家畜によって生活する。
9	Сар гарав.	月は出た.	月が出た。
10	Од ч гарав.	スターも出た.	星も出た。
11	Үхэр мөөрнө.	牛がモーと鳴く.	牛がモーと鳴く。
12	Хонь майлна.	羊がメーと鳴く.	羊がメーと鳴く。
13	Эгч дээл хувцас оёж байна.	姉服を服を縫っている.	姉が着物を縫っています。
14	Хавар болов.	春になった.	春になった。
15	Шувуу донгодно.	鳥がさえずる.	鳥が囀ります。
16	Тунгалаг гол урсаж байна.	清らかな河が流れている.	きれいな河が流れています。
17	Нүүр, гараа сайн угаа!	顔を，スタートがよく洗いなさい！	顔と手をよく洗いなさい。
18	Уул тал цасаар хучигдав.	山草原で白くなりながらおおわれた.	山野が雪で蔽われた。
19	Би үзгээр үсэг бичнэ.	私はペンで字を書く.	私はペンで文字を書く。
20	Дорноговь аймгийн Дэлгэрхэн сумын нутагт чоно ховор учир, энэ сумын хэдэн анчин Хэнтий, Сүхбаатар аймгийн нутаг руу тэмээгээр явж чоно авлав.	ドルノゴビ県のデルゲルヘン郡の地方に狼が珍しい理由に，この郡の若干猟師ヘンティ，スフバートル県の土地の方へらくだで行って狼が狩った.	ドルノゴビ県のデルゲルヘン郡の遊牧地には、狼が少ないので、この郡の数人の猟師は、ヘンティー、スフバートル県の遊牧地の方へ駱駝で行って、狼を狩った。